Factorized Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures

Tomi Silander, Teemu Roos, Petri Kontkanen and Petri Myllymäki
Helsinki Institute for Information Technology HIIT, Finland

Abstract

This paper introduces a new scoring criterion, factorized normalized maximum likelihood, for learning Bayesian network structures. The proposed scoring criterion requires no parameter tuning, and it is decomposable and asymptotically consistent. We compare the new scoring criterion to other scoring criteria and describe its practical implementation. Empirical tests confirm its good performance.

1 Introduction

The popular Bayesian criterion, BDeu (Buntine, 1991), for learning Bayesian network structures has recently been reported to be very sensitive to the choice of prior hyperparameters (Silander et al., 2007). On the other hand, general model selection criteria, such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), are derived through asymptotics and their behavior is suboptimal for small sample sizes. The study of different scoring criteria is further complicated by the fact that learning the network structure is NP-hard for all popular scoring criteria (Chickering, 1996), even if these criteria have a convenient characteristic of decomposability, which allows incremental scoring in heuristic local search (Heckerman et al., 1995). Due to recent advances in exact structure learning (Koivisto and Sood, 2004; Silander and Myllymäki, 2006) it is feasible to find the optimal network for decomposable scores when the number of variables is less than about 30. This makes it possible to study the behavior of different scoring criteria without the uncertainty stemming from the heuristic search.

In this paper we introduce a new decomposable scoring criterion for learning Bayesian network structures, the *factorized normalized maximum likelihood* (fNML). This score features no tunable parameters, thus avoiding the sensitivity problems of Bayesian scores. We show that the new criterion is asymptotically consistent. Unlike AIC and BIC, it is derived based on optimality criterion for finite sample sizes, and it has a probabilistic interpretation.

The rest of the paper is structured as follows. In Section 2, we will first introduce Bayesian networks and the notation needed later. In Section 3, we review the most popular decomposable scores, after which in Section 4, we are ready to introduce the fNML criterion. We then briefly discuss the implementation of this new score in Section 5. Section 6 presents the empirical experiments, and the conclusions are summarized in Section 7.

2 Bayesian Networks

We assume that reader is familiar with Bayesian networks (for tutorial, see (Heckerman, 1996)), and only introduce the notation needed later in this paper.

A Bayesian network defines a joint probability distribution for an m-dimensional multivariate data vector $X = (X_1, \ldots, X_m)$. We assume that all variables are discrete, so that variable X_i may have r_i different values $\{1, \ldots, r_i\}$.

A Bayesian network consists of a directed acyclic graph G and a set of conditional probability distributions. We specify the DAG with a vector $G = (G_1, \ldots, G_m)$ of parent sets so that $G_i \subset \{X_1, \ldots, X_m\}$ denotes the parents of variable X_i, i.e., the variables from which there is an arc to X_i. Each parent set G_i has q_i ($q_i = \prod_{X_p \in G_i} r_p$) possible values that are the possible value combinations of the variables belonging to G_i. We assume a non-ambiguous enumeration of these values and denote the fact that G_i holds the j^{th} value combination simply by $G_i = j$.

The local Markov property for Bayesian networks states that each variable is independent of its non-descendants given its parents. Functionally this is equivalent to the following factorization of the joint distribution

$$P(x \mid G) = \prod_{i=1}^{m} P(x_i \mid G_i). \tag{1}$$

The conditional probability distributions $P(X_i \mid G_i)$ are determined by a set of parameters, Θ, via the equation

$$P(X_i = k \mid G_i = j, \Theta) = \theta_{ijk},$$

where k is a value of X_i, and j is a value configuration of the parent set G_i. We denote the set of parameters associated with variable X_i by Θ_i.

For learning Bayesian network structures we assume a data D of N complete i.i.d instantiations of the vector X, i.e., an $N \times m$ data matrix without missing values. It turns out to be useful to introduce a notation for certain parts of this data matrix. We often want to select rows of the data matrix by certain criteria. We then write the selection criterion as a superscript of the data matrix D. For example, $D^{G_i=j}$ denotes those rows of D where the variables of G_i have the j^{th} value combination. If we further want to select certain columns of these rows, we denote the columns by subscripting D with a corresponding variable set. As a shorthand, we write $D_{\{X_i\}} = D_i$. For example, $D_i^{G_i=j}$ selects the i^{th} column of the rows $D^{G_i=j}$.

Since the rows of D are assumed to be i.i.d, the probability of a data matrix can be calculated just by taking the product of the row probabilities. Combining equal terms yields

$$P(D \mid G, \Theta) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}, \tag{2}$$

where N_{ijk} denotes number of rows in $D^{X_i=k,G_i=j}$.

For a given structure G, we use notation $\hat{P}(D \mid G) = \sup_\theta P(D \mid G, \theta)$. The maximizing parameters are simply the relative frequencies found in data: $\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}}$, where N_{ij} denotes the number of rows in $D^{G_i=j}$, or 1.0 if $N_{ij} = 0$. We often drop the dependency on G when it is clear from the context.

3 Decomposable scores

In general, a scoring function $\textsc{Score}(G, D)$ for learning a Bayesian network structure is called decomposable, if it can be expressed as a sum of local scores

$$\textsc{Score}(G, D) = \sum_{i=1}^{m} S(D_i, D_{G_i}). \tag{3}$$

Many popular scoring functions avoid overfitting by balancing the fit to the data with the complexity of the model. A common form of this idea can be expressed as

$$\textsc{Score}(G, D) = \log \hat{P}(D \mid G) - \Delta(D, G), \tag{4}$$

where $\Delta(D, G)$ is a complexity penalty.

The maximized likelihood $\hat{P}(D \mid G)$ decomposes by the network structure, and for the decomposable scores handled in this paper, the complexity penalty decomposes too. Hence, we can write the penalized scores in the decomposed form (3), with the local scores given by

$$S(D_i, D_{G_i}) = \log \hat{P}(D_i \mid D_{G_i}) + \Delta_i(D_i, D_{G_i}). \tag{5}$$

Different scores differ in how the local penalty $\Delta_i(D_i, D_{G_i})$ is determined.

3.1 AIC and BIC

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two popular decomposable scores for learning Bayesian network structures. These scores do not have any additional parameters so in this sense they are similar to the proposed fNML score. The penalty terms for these scores are $\Delta_i^{\text{BIC}} = \frac{q_i(r_i-1)}{2} \ln N$, and $\Delta_i^{\text{AIC}} = q_i(r_i - 1)$. Both of these complexities are independent of the data, and only depend on the arities r_i of random variables and the structure of the Bayesian network.

3.2 Bayesian Dirichlet scores

Bayesian Dirichlet (BD) scores assume that the parameter vectors Θ_{ij} are independent of each other and distributed by Dirichlet distributions with hyper-parameter vector $\vec{\alpha}_{ij}$. Given a vector of hyper-parameters $\vec{\alpha}$, the local score can be written as

$$\begin{aligned} S_{\mathrm{BD}}(D_i, D_{G_i}, \vec{\alpha}) &= \log P(D_i \mid D_{G_i}, \vec{\alpha}) \\ &= \sum_{j=1}^{q_i} \log P(D_i^{G_i=j} \mid D_{G_i}^{G_i=j}, \vec{\alpha}_{ij}) \\ &= \sum_{j=1}^{q_i} \log \left(\frac{B(\vec{\alpha}_{ij} + \vec{N}_{ij})}{B(\vec{\alpha}_{ij})} \right), \end{aligned}$$

where B is a multinomial Beta function

$$B(\alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{i=1}^{K} \alpha_k)}.$$

With all $\alpha_{ijk} = 1$ we get a K2-score (Cooper and Herskovits, 1992), and with $\alpha_{ijk} = \frac{\alpha}{q_i r_i}$ we get a family of BDeu scores popular for giving equal scores to different Bayesian network structures that encode the same independence assumptions. BDeu scores depend only on a single parameter, the *equivalent sample size* α. Recent studies on the role of this parameter show that network learning under BDeu is very sensitive to this parameter (Silander et al., 2007).

For comparison, we can write the BD-score as a penalized maximized likelihood with penalty

$$\Delta_i^{BD}(D_i, D_{G_i}) = \sum_{i=i}^{q_i} \log \left(\frac{\hat{P}(D_i^{G_i=j} \mid D_{G_i}^{G_i=j})}{P(D_i^{G_i=j} \mid D_{G_i}^{G_i=j}, \vec{\alpha}_{ij})} \right). \tag{6}$$

We immediately notice that this penalty is always positive. The complexity is data-dependent and it is controlled by the hyper-parameters α_{ijk}. The asymptotic behavior of this Bayesian regret is well studied (Grünwald, 2007). However, when learning Bayesian networks, the data parts $D_i^{G_i=j}$ are often very small, which makes asymptotic result less informative.

4 fNML

The *factorized normalized maximum likelihood* (fNML) score is based on the *normalized maximum likelihood* (NML) distribution (Shtarkov, 1987; Rissanen, 1996). The NML distribution for the model class $\mathcal{M}$ (which may or may not be a Bayesian network) is the unique distribution solving the minimax problem

$$\min_Q \max_{D'} \frac{\hat{P}(D' \mid \mathcal{M})}{Q(D' \mid \mathcal{M})}, \tag{7}$$

where Q ranges over all distributions.

As originally shown by Shtarkov (1987) the solution of the above minimax problem is given by

$$P_{\mathrm{NML}}(D \mid \mathcal{M}) = \frac{\hat{P}(D \mid \mathcal{M})}{\sum_{D'} \hat{P}(D' \mid \mathcal{M})}, \tag{8}$$

where the normalization is over all data sets D' of a fixed size N. The log of the normalizing factor is called *parametric complexity* or *regret.*

Evaluation of the normalizing sum is often hard due to exponential number of terms in the sum. Currently, there are tractable formulas for only a handful of models; for examples, see (Grünwald, 2007). In the case of a single r-ary multinomial variable and the sample size n the normalizing sum is given by

$$\mathcal{C}_n^r = \sum_{k_1+k_2+\ldots+k_r=n} \frac{n!}{k_1!\,k_2!\cdots k_r!} \prod_{j=1}^{r} \left(\frac{k_j}{n}\right)^{k_j}, \tag{9}$$

where the sum goes over all non-negative integer vectors $(k_j)_{j=1}^r$ that sum to n. A linear-time algorithm for the computation of $\mathcal{C}_n^r$ was introduced recently by Kontkanen and Myllymäki (2007).

Given a data set D, the NML model selection criterion proposes to choose the model $\mathcal{M}$ for which the $P_{\mathrm{NML}}(D \mid \mathcal{M})$ is largest. After taking the logarithm the score is in a form of penalized log likelihood with complexity penalty describing how well the model can fit any equal size dataset D'.

Because of the score equivalence of the maximum likelihood score, the NML score is score equivalent as well. However, it is not decomposable, and the parent assignment problem is known to be NP-hard (Koivisto, 2006). Sacrificing the score equivalence we propose a decomposable version of this score, which penalizes the complexity locally similarly to the other decomposable scores. Specifically, we propose the local score

$$\begin{aligned} S_{\mathrm{fNML}}(D_i, D_{G_i}) &= \log P_{\mathrm{NML}}(D_i \mid D_{G_i}) \qquad (10) \\ &= \log \left(\frac{\hat{P}(D_i \mid D_{G_i})}{\sum_{D'_i} \hat{P}(D'_i \mid D_{G_i})} \right), \end{aligned}$$

where the normalizing sum goes over all the possible D_i-column vectors of length N, i.e., $D'_i \in \{1, \ldots, r_i\}^N$.

Since equation (10) defines a (log) conditional distribution for the data column D_i, adding these local scores together yields a total score that defines a distribution for the whole data. In this sense fNML can be seen as an alternative way to define the marginal likelihood for the data

$$\log P_{\mathrm{fNML}}(D \mid G) = \sum_{i=1}^{m} \log P_{\mathrm{NML}}(D_i \mid D_{G_i}).$$

At the same time, combining the local scores yields an enumerator that equals the decomposition of the maximum likelihood, thus the whole score can be seen as a penalized maximum log-likelihood with local (data-dependent) penalties

$$\Delta_i^{\mathrm{fNML}}(D_{G_i}) = \log \sum_{D'_i} \hat{P}(D'_i \mid D_{G_i}). \qquad (11)$$

The following observation follows from the factorization of the maximum likelihood by the parent configurations, and it is crucial for efficient calculation of the local penalty term.

Theorem 1. *The local penalty of fNML can be expressed in terms of multinomial normalizing constants*

$$\Delta_i^{\mathrm{fNML}}(D_{G_i}) = \sum_{j=1}^{q_i} \log \mathcal{C}_{N_{ij}}^{r_i},$$

where $\mathcal{C}_{N_{ij}}^{r_i}$ is the normalizing constant of NML for an r_i-ary multinomial model with sample size N_{ij}.

The theorem follows by noting that the maximized likelihood $\hat{P}(D_i \mid D_{G_i})$ factorizes into independent parts according to the values of D_{G_i}.

To conclude this section we show that asymptotically, and under mild regularity conditions, the fNML score belongs to the (large) class of BIC-like scores that are consistent. Other scores in this class include most Bayesian and MDL criteria. The regularity conditions required for BIC-like behavior typically exempt a measure zero set of generating parameters, such as the boundaries of the parameter simplex. The following theorem gives sufficient conditions on the penalty term that guarantee consistency for exponential family models.

Theorem 2 (Remark 1.2 in (Haughton, 1988))**.** *For (curved) exponential families, if data is generated by an i.i.d. distribution p, and the penalty term is given by $\frac{1}{2} k\, a_N$, where k is the number of parameters and a_N is a sequence of positive real numbers, satisfying*

$$a_N/N \to 0, \quad \textit{and} \quad a_N \to \infty,$$

as $N \to \infty$, then, symptotically, the model containing p that has the least number of parameters will be chosen.

Since Bayesian networks are curved exponential families (Geiger et al., 2001; Chickering, 2002), it now remains to prove that the penalty term of fNML satisfies this property.

Theorem 3 (Asymptotically fNML behaves like BIC)**.** *Assuming that the maximum likelihood parameters are asymptotically bounded away from the boundaries of the parameter simplex, the local penalty of fNML behaves as*

$$\Delta_i^{\mathrm{fNML}}(D_{G_i}) = \frac{q_i(r_i - 1)}{2} \log N + \mathcal{O}(1),$$

almost surely, where the $\mathcal{O}(1)$ term is bounded by a constant wrt. N.

Proof. By Thm. 1, the local penalty is a sum of logarithms of multinomial normalizing constants. The latter is known to grow as $\log \mathcal{C}_{N_{ij}}^{r_i} = \frac{r_i-1}{2} \log N_{ij} + \mathcal{O}(1)$, (Rissanen, 1996). Under the assumption that the maximum likelihood parameters are bounded away from the boundaries, the counts N_{ij} grow linearly in the total sample size N almost surely, which implies that we have $\log N_{ij} = \log([\eta + o(1)]N) = \log N + \mathcal{O}(1)$ with some $0 < \eta < 1$. Adding together the q_i terms yields the result. □

Since $q_i(r_i - 1)$ is the number of parameters (associated with the ith variable), the property of Thm. 2 holds for the fNML penalty.

5 Implementation

We now provide information for practical implementation of the fNML score for Bayesian networks. Due to the decomposability of the score the only new implementational issue for the fNML is to calculate the terms $\mathcal{C}_n^r$ of the Thm. 1. For reasonable N and R ($R = \max r_i$) these values can be stored in an $N \times R$ table, which can be done before structure learning. Moreover, this table does not depend on data or any parameters, so it can be done just once.

The calculation of the $\mathcal{C}$-table with N rows and R columns proceeds as follows. First of all, $\mathcal{C}_0^r = 1$ for all r, and $\mathcal{C}_n^1 = 1$ for all n. For $r = 2$ we can use the formula (9), which yields

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{N}\right)^{n-h}, \qquad (12)$$

and for $r > 2$ we can use the recursion (Kontkanen and Myllymäki, 2007)

$$\mathcal{C}_n^r = \mathcal{C}_n^{r-1} + \frac{n}{r-2}\mathcal{C}_n^{r-2}. \qquad (13)$$

Calculating the column $\mathcal{C}_*^2$ using the formula (12) takes time $\mathcal{O}(N^2)$, and the calculation of the rest of the table using the formula (13) takes just $\mathcal{O}(NK)$. For very large N, the complexity of calculating the column $\mathcal{C}_*^2$ may be prohibitive. In this case a very accurate Szpankowski approximation (Kontkanen et al., 2003)

$$\mathcal{C}_n^2 = \frac{n\pi}{2} e^{\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}} \qquad (14)$$

can be used.

If the space for storing the table is critical, one may just store 1000 first entries of column $\mathcal{C}_*^2$, use Szpankowski approximation for the rest of the column, and use formula (13) for calculating the values for $r > 2$.

6 Experiments

It is not obvious how to compare different criteria for learning Bayesian network structures. If the data is generated from a Bayesian network, one might call for selecting the data generating network, but if the generating network is complex, and the sample size is small, it may be rational to pick a simpler model.

This simplicity requirement is often backed up by arguments about the generalization capability of the model. However, it is not always clear how the network structure should be used for prediction.

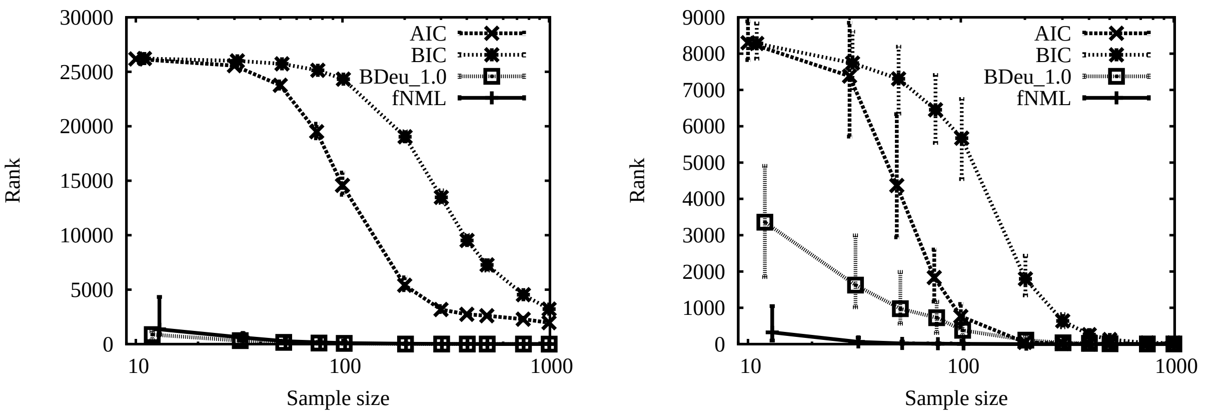
A softer version of discovering the generating model is to compute a structural distance measure between the selected and the generating network structures. A common choice is to calculate an editing distance with operations such as arc additions, deletions and reversals. Even if we take the generating structure as a golden standard, this approach is problematic, since these editing operations are not independent. For example, fixing a certain arc can lead to several other changes to the network structure if the selection by a score is made only among the structures having the fixed arc present.

Despite of these problems in the empirical testing, we conducted a golden standard experiment. We first generated data from different networks with five nodes, and then studied how the generating network structures were ranked among all the possible networks by different scoring criteria.

For BDeu and fNML scores that both calculate the probability $P(D \mid G)$, we also compared the scores for the real data sets. This experiment can be seen as the result of a sequential prediction competition, since by the chain rule we can write

$$P(D \mid G) = \prod_{i=1}^{N} P(d_i \mid G, d^{i-1}), \qquad (15)$$

where d_i is the ith data vector, and $d^{i-1} = \{d_1, \ldots, d_{i-1}\}$ denotes the first $i-1$ vectors. The idea follows the principle of prequential model selection (Dawid, 1984).



(a) BDeu Scheme

(b) Dir(1/2,. . . ,1/2) Scheme

Figure 1: The median curves for different scoring criteria as a function of sample size when the parameters for a 5-node, 7-edge network were generated by the BDeu and Dir(1/2,. . . ,1/2) schemes. Errorbars indicate upper and lower quartiles.

We will now explain the experiments in more detail.

6.1 Artificial data

We first compared the ability of different scoring criteria to discover the data generating structure. For this purpose we generated 100 different 5-node Bayesian network structures with 4 edges and another 100 structures with 7 edges. The variables were randomly assigned to have 2 – 4 values ($r_i \in \{2, 3, 4\}$). For each network, we generated parameters by two different schemes. The first scheme exactly matched the assumptions of the BDeu score with $\alpha = 1$, i.e., the parameters were distributed by $\theta_{ij} \sim Dir(\frac{1}{r_i q_j}, \ldots, \frac{1}{r_i q_j})$. The other scheme was to generate the parameters independently from a Dirichlet distribution $\theta_{ij} \sim Dir(1/2, \ldots, 1/2)$. This distribution was selected instead of the uniform distribution in order to make the generating structure more identifiable.

For each network (structure + parameters), we generated 100 data sets of 1000 data vectors, and studied how different scoring criteria ranked the structure of the generating network among all the 5-node networks as a function of (sub)sample size.

Not surprisingly, the results indicate that when parameter generation mechanism matches the assumptions of the BDeu-score, the BDeu usually also ranks the generating structure higher than the other scores (Figure 1(a)). However, fNML usually behaves very similarly to BDeu. The density of the network (4 vs. 7 edges) is not a very significant factor. If anything, the similar behavior of fNML and BDeu is more pronounced in networks with 7 edges. For the parameter-free scores, AIC and BIC, the underfitting tendency of BIC can be clearly detected whereas AIC tends to rank the generating network higher. Qualitatively these two scores seem to behave similarly to each other.

Switching the parameter generation scheme to independent Dirichlets with $\alpha_{ijk} = 0.5$ usually also switches the ranking ability of fNML and BDeu, while the behavior of AIC and BIC stays mostly unaffected. For example, Figure 1(b) was generated using the same network structure as for Figure 1(a). Only the parameter generation scheme was changed from BDeu to Dir. For dense networks fNML often appears as a clear winner.

6.2 Real data

Learning the structure with AIC or BIC does not readily suggest any particular way to use the learned structure for prediction, but the prequential interpretation of the BDeu score and the fNML allows comparison. However, the BDeu score is known to be very sensitive to the

Table 1: Summary of the prediction experiment.

Data	N	m	#vals	α^*	BDeu1	BDeu*	fNML
balance	625	5	4.6	48	-4549.06	**-4445.64**	-4478.36
iris	150	5	3.0	2	-452.21	**-449.71**	-450.90
thyroid	215	6	3.0	2	-577.52	-575.55	**-572.42**
liver	345	7	2.9	4	-1309.67	-1299.83	**-1299.38**
ecoli	336	8	3.4	8	-1715.92	-1661.34	**-1643.64**
abalone	4177	9	3.0	6	-15946.58	-15891.25	**-15847.33**
diabetes	768	9	2.9	4	-3678.57	-3662.31	**-3654.02**
post operative	90	9	2.9	3	-647.35	-642.98	**-639.94**
yeast	1484	9	3.7	6	-7938.60	-7873.21	**-7848.98**
breast cancer	286	10	4.3	8	-2781.62	**-2737.20**	-2739.34
shuttle	58000	10	3.0	3	*-97635.72*	**-97620.78**	-97714.22
tic tac toe	958	10	2.9	51	-9423.07	**-9126.78**	-9162.39
bc wisconsin	699	11	2.9	8	-3315.51	-3262.33	**-3239.56**
glass	214	11	3.3	6	-1288.93	-1255.73	**-1233.18**
page blocks	5473	11	3.2	3	-12455.60	-12438.01	**-12410.69**
heart cleveland	303	14	3.1	13	-3450.07	-3356.78	**-3352.32**
heart hungarian	294	14	2.6	5	-2376.53	-2348.23	**-2343.65**
heart statlog	270	14	2.9	10	-2867.54	-2819.37	**-2814.28**
wine	178	14	3.0	8	-1866.41	-1821.28	**-1808.66**
adult	32561	15	7.9	50	-329373.73	-326803.91	**-326486.85**

equivalent sample size parameter, which creates an extra complication.

For predictive comparison we selected 20 UCI data sets[1] for which the score maximizing hyperparameter α has been reported (Silander et al., 2007), and we compared the maximum fNML scores to the maximum scores obtained with BDeu1 (BDeu with $\alpha = 1.0$) and BDeu* (BDeu with score maximizing α). In reality, we do not know the score maximizing α's, and searching structures with many α is usually computationally too hard. Optimal structures were obtained by the exact structure learning algorithm described in (Silander and Myllymäki, 2006).

Table 1 lists for each data set the number of data vectors N, the number of variables m, the average number of values per variable #vals, the BDeu maximizing equivalent sample size parameter $\alpha*$ (with integer precision), and the actual scores obtained with three different scoring criteria. The score obtained with fNML is the best of the three 14 times out of 20, and only once BDeu1 yields higher score than fNML.

7 Conclusions

We have introduced a new probabilistic scoring criterion, the factorized normalized maximum likelihood, for learning Bayesian network structures from complete discrete data. The score aims at being an efficient and parameter-free criterion for finite sample sizes. The score is also decomposable, which makes it possible to use it with existing search heuristics and exact structure learning algorithms.

Initial empirical tests are promising. We are particularly pleased with fNML's ability to learn network structures with good predictive capabilities. While lot more empirical work has to be done, the current experiments already show a great promise for a good and care free

[1] `http://www.ics.uci.edu/~mlearn/MLRepository.html`

scoring criterion for learning Bayesian network structures.

Acknowledgments

The authors thank the anonymous referees for valuable comments. This work was supported in part by the Academy of Finland (Project Civi), by the Finnish Funding Agency for Technology and Innovation (Kukot, PMMA), and the IST Programme of the European Community, under the PASCAL Network of Excellence.

References

H. Akaike. 1973. Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, editors, *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai Kiado.

W. Buntine. 1991. Theory refinement on Bayesian networks. In B. D'Ambrosio, P. Smets, and P. Bonissone, editors, *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers.

D.M. Chickering. 1996. Learning Bayesian networks is NP-Complete. In D. Fisher and H. Lenz, editors, *Learning from Data: Artificial Intelligence and Statistics V*, pages 121–130. Springer-Verlag.

D.M. Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

G. Cooper and E. Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

A.P. Dawid. 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, 147:278–292.

D. Geiger, D. Heckerman, H. King, and C. Meek. 2001. Stratified exponential families: graphical models and model selection. *Annals of Statistics*, 29:505–529.

P. Grünwald. 2007. *The Minimum Description Length Principle*. MIT Press.

D.M.A. Haughton. 1988. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355.

D. Heckerman, D. Geiger, and D.M. Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, September.

D. Heckerman. 1996. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, One Microsoft Way, Redmond, WA 98052.

M. Koivisto and K. Sood. 2004. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, May.

M. Koivisto. 2006. Parent assignment is hard for the MDL, AIC, and NML costs. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT-06)*, pages 289–303.

P. Kontkanen and P. Myllymäki. 2007. A linear-time algorithm for computing the multinomial stochastic complexity. *Information Processing Letters*, 103(6):227–233.

P. Kontkanen, W. Buntine, P. Myllymäki, J. Rissanen, and H. Tirri. 2003. Efficient computation of stochastic complexity. In C. Bishop and B. Frey, editors, *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, pages 233–238. Society for Artificial Intelligence and Statistics.

J. Rissanen. 1996. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January.

G. Schwarz. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Yu.M. Shtarkov. 1987. Universal sequential coding of single messages. *Problems of Information Transmission*, 23:3–17.

T. Silander and P. Myllymäki. 2006. A simple approach for finding the globally optimal Bayesian network structure. In R. Dechter and T. Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 445–452. AUAI Press.

T. Silander, P. Kontkanen, and P. Myllymäki. 2007. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In R. Parr and L. van der Gaag, editors, *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360–367. AUAI Press.